Multimodal image retrieval

Fusing modalities with multilayer multimodal pLSA

Stefan Romberg · Rainer Lienhart · Eva Hörster

Abstract In this work, we extend the standard single-layer *probabilistic Latent Semantic Analysis* (*pLSA*) (Hofmann in Mach Learn 42(1–2):177–196, 2001) to multiple layers. As multiple layers should naturally handle multiple modalities and a hierarchy of abstractions, we denote this new approach *multilayer multimodal probabilistic Latent Semantic Analysis* (*mm-pLSA*). We derive the training and inference rules for the smallest possible non-degenerated mm-pLSA model: a model with two leaf-pLSAs and a single top-level pLSA node merging the two leaf-pLSAs. We evaluate this approach on two pairs of different modalities: SIFT features and image annotations (tags) as well as the combination of SIFT and HOG features. We also propose a fast and strictly stepwise forward procedure to initialize the bottom–up mm-pLSA model, which in turn can then be post-optimized by the general mm-pLSA learning algorithm. The proposed approach is evaluated in a query-by-example retrieval task where various variants of our mm-pLSA system are compared to systems relying on a single modality and other ad-hoc combinations of feature histograms. We further describe possible pitfalls of the mm-pLSA training and analyze the resulting model yielding an intuitive explanation of its behaviour.

Keywords Topic models · Image retrieval · Hierarchical pLSA · pLSA · SIFT · HOG · Image annotation

S. Romberg (✉) · R. Lienhart · E. Hörster
Multimedia Computing and Computer Vision Lab,
University of Augsburg, Augsburg, Germany
e-mail: Stefan.romberg@informatik.uni-augsburg.de

R. Lienhart
e-mail: lienhart@informatik.uni-augsburg.de

1 Introduction

Many content-based image retrieval systems either solely rely on visual features or on text features to derive a representation of the image content. This is especially true for systems using topic models based on *probabilistic Latent Semantic Analysis* (*pLSA*) [7,16,22]. There are good reasons why pLSA is applied to unimodal data: The straightforward application of pLSA to multimodal data by subsuming all words of the various modes (which are generally derived from appropriate features of the respective modality) into one large word set (called vocabulary) frequently does not lead to the expected improvement in retrieval performance. Even mixing words derived from different kinds of features within one domain such as different kinds of visual salient point descriptors (e.g., SIFT [23], SURF [2], Geometric blur [3], or self-similarity feature [28]) using different sampling strategies (e.g., dense versus sparse sampling) does not work satisfactorily with this obvious application of pLSA.

Thus, we propose a multilayer multimodal pLSA model (referred to as *mm-pLSA*) that can handle different modalities as well as different features within a mode effectively and efficiently. This model utilizes not just a single layer of topics or aspects, but a hierarchy of topics. We introduce the overall approach by using the smallest possible non-degenerated mm-pLSA model: a model with two separate sets of (leaf-)topics for data from two different modes and a set of top-level topics that merges the knowledge of the two sets of leaf-topics. This approach resembles somewhat the computation of two independent leaf-pLSAs from two different data modalities, whose topics in turn are merged by a single top-level pLSA node, and thus lends the proposed approach its name: *mm-pLSA*. From this derivation, it is obvious how to extend the learning and inference rules to more modalities and more layers. We also propose a fast and strictly

stepwise forward procedure to initialize the bottom–up mm-pLSA model that leads to much better learning results of the mm-pLSA learning algorithm compared to random initialization.

The paper is organized as follows. Section 2 summarizes related work. In Sect. 3, we first describe the model of the standard pLSA algorithm (Sect. 3.1) as well as how to learn a pLSA model in general (Sect. 3.2) and specifically from the visual features (Sect. 3.3) and tag features (Sect. 3.5). Classification of a new image or text document is also addressed. Then, Sect. 4 presents the core novelty of our work in detail: the *multilayer multimodal probabilistic Latent Semantic Analysis* model (*mm-pLSA*). It starts in Sect. 4.1 with a motivation and a detailed explanation of the model, before we derive the training and inference steps in Sect. 4.2. A heuristic for fast and good initialization of the multilayer multimodal pLSA model is presented in Sect. 4.3 and carefully evaluated in Sect. 5 on a large-scale database consisting of 10 million images downloaded from Flickr. Our proposed mm-pLSA-based image retrieval system is compared to systems relying solely on visual features [22] or tag features as well as to a pLSA-based system with the combined vocabulary set from the visual and tag domain. Moreover, we compare the mm-pLSA based image retrieval system on multiple, same domain features to systems based on a single feature and other ad-hoc combinations of these. In addition, further insights of the resulting model are presented before Sect. 6 concludes the paper.

2 Related work

Topic models have been used in several previous works to derive a low-dimensional image description suitable for large-scale image retrieval. For example, [22] uses probabilistic Latent Semantic Analysis (pLSA [16]) based models, [18] applies Latent Dirichlet Allocation (LDA [6]) to derive a topic representation, and [13] adopts the Correlated Topic Model (CTM [4]). However, all of the previous mentioned works build their image representation solely on visual features.

In [1,5,24], the authors propose topic models to model annotated image databases. They use the models to automatically annotate images and/or image regions. One key difference of our work to those previous works is that we build an image retrieval system instead of annotating images. Moreover, the image database we use for learning and retrieval is a real-world, large-scale, 10 million images' database in contrast to the small and almost noise-free COREL data-base that was used in the above works for learning and testing. Thus, in our case the tags associated with an image do not necessarily refer to the visual content shown. For example, they may also denote the time, date, place, or circumstances under which the picture was taken. This makes models, which try to associate image regions directly with tags, difficult to learn and apply.

Our approach uses a hierarchical model as we have more than one topic layer. In [29], the authors adapt the Hierarchical Latent Dirichlet Allocation (hLDA) model, which has been developed originally for the unsupervised discovery of topic hierarchies in text, to the visual domain. They use the model for object classification and segmentation. However their model only accounts for one modality: visual features. Moreover, appropriate initialization of the complex model is difficult. Another example of a hierarchical model for image content are deep networks [15,17] with which—on a very high-level point of view—we share the stepwise forward initialization and subsequent optimization.

The multi-feature pLSA [32] is somewhat similar to our approach, but uses only a single topic layer that models the co-occurrence of visual features of two different types at once.

This article is a substantial extension of our previous published work [21], which much more thoroughly analysis the strengths and weaknesses of our proposed mm-pLSA model.

3 Standard pLSA

3.1 Motivation and model

The pLSA was originally devised by Hofmann [16] in the context of text document retrieval, where words constitute the elementary parts of documents. Applied to images, each image represents a single visual document. pLSA can be applied directly to image tags, as tags are simply words. However, for our visual features we need comparable elementary parts called visual words. For the moment we assume that all features we computed in a given mode are somehow mapped to words in that mode. Details of the mapping from the visual features to the mode-specific words are given in Sect. 3.3. For now we just assume that we have words.

The key concept of the pLSA model is to map the high-dimensional word distribution vector of a document to a lower dimensional *topic vector* (also called *aspect vector*). Therefore, pLSA introduces a latent, i.e. unobservable topic layer between the documents (i.e. images here) and the observed words. It is assumed that each document consists of a mixture of multiple topics and that the occurrences of words (i.e., visual words in the images or tags of images, respectively) is a result of the topic mixture. This generative model is expressed by the following probabilistic model:

$$P(d_i, w_j) = P(d_i) \sum_K P(z_k|d_i) P(w_j|z_k) \tag{1}$$

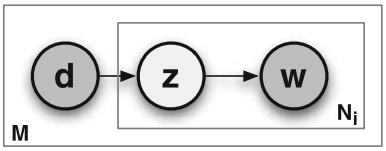

Fig. 1 Standard pLSA-model

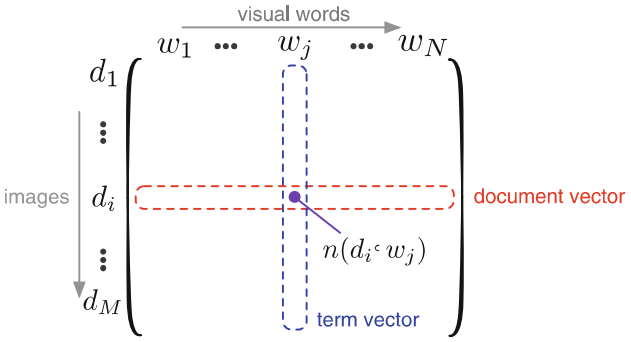

Fig. 2 Term-document matrix

where $P(d_i)$ denotes the probability of a document d_i of the database to be picked, $P(z_k|d_i)$ the probability of a topic z_k given the current document, and $P(w_j|z_k)$ the probability of a visual word w_j given a topic. The model is graphically depicted in Fig. 1. N_i denotes the number of words of which document d_i consists. In total we assume M documents. It is important not to confuse N_i, the number of words in document d_i, with N, the number of words in the vocabulary.

Once a topic mixture $P(z_k|d_i)$ is derived for each document d_i, a high-level representation has been found based on the respective mode to which the words belong. At the same time, this representation is of low dimensionality as we commonly choose the number of concepts in our model to be much smaller than the number of words. The K-dimensional topic vector can be used directly in a query-by-example retrieval task, if we measure document similarity by computing the L_1, L_2, or cosine distance between topic vectors of different documents.

3.2 Training and inference

Computing a *term-document matrix* of the training corpus is a prerequisite for deriving a pLSA model (see Fig. 2). Each entry in row i and column j of the term-document matrix $[n(d_i, w_j)]_{i,j}$ specifies the absolute count with which word w_j (also called a term) occurs in document d_i. The terms are taken from a predefined dictionary consisting of N terms. The number of documents is M. Note that by normalizing each document vector to 1 using the L1-norm, the document vector $(n(d_i, w_1), \ldots, n(d_i, w_N))$ of d_i becomes the estimated mass probability distribution $P(w_j|d_i)$.

We learn the unobservable probability distributions $P(z_k|d_i)$ and $P(w_j|z_k)$ from the observable data $P(w_j|d_i)$ and $P(d_i)$ using the Expectation-Maximization algorithm (EM-Algorithm) [8,16]:

E-Step:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \tag{2}$$

M-Step:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j=1}^{N}\sum_{i=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)} \tag{3}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{n(d_i)} \tag{4}$$

Given a new test image d_{test}, we estimate the topic probabilities $P(z_k|d_{\text{test}})$ from the observed words. The sole difference between inference and learning is that the K learned conditional word distributions $P(w_j|z_k)$ are never updated during inference. Thus, only Eqs. (2) and (4) are iteratively updated during inference.

3.3 Visual pLSA-model

The first step in building a bag-of-words representation for the visual content of images is to extract visual features from each image. In our case, we apply dense sampling with a vertical and horizontal step size of 10 pixels across the image pyramid created with a scale factor of $1/\sqrt[4]{2}$ in order to extract local image features at regular grid points. SIFT descriptors [23] computed over a local region of 41×41 pixels are used to describe the grayscale image regions around each grid point in an orientation invariant fashion. Although we use SIFT features in this work, any other feature could be used instead.

Next, the 128-dimensional real-valued local image features have to be quantized into discrete visual words to derive a finite vocabulary. Quantization of the features into visual words is performed using a flat vocabulary derived by k-means clustering [30]. In contrast to our previous work we use a flat vocabulary rather than a vocabulary tree [25] as the hierarchical k-means clustering of the feature space has been shown to be inferior to standard or approximate k-means in previous works [26]. Also, speed is not a big issue with a vocabulary size of 10,000 visual words, which we will use in our experiments.

Once a visual vocabulary of size N^v is determined, we map all descriptor vectors of an image to their closest visual words and build the document vector that holds the counts of the visual word occurrences in the corresponding image by incrementing the associated word count. Note that this very popular image description does not preserve any spatial relationship between the occurrences of the visual words.

The image is simply modeled as a histogram (bag) of its visual words.

The document vectors (also called *co-occurrence vectors*) of randomly selected training images are then used to train a pLSA model. Once a pLSA model is learned, it can be applied to all images in the database and hence derive a vector representation for each image, where the vector elements denote the degree to which an image depicts a certain visual topic. Given a query image and its topic distribution the retrieval then works by finding the top r images with the closest topic distribution to the query topic distribution in the database.

3.4 Fusion of multiple visual features

In this work, we also evaluate how the proposed multilayer multimodal approach is able to combine different visual features. In this particular case, we use the mm-pLSA to combine SIFT and HOG features.

The basis for our 2×2 HOG features are the improved, 31-dimensional HOG cell features of [12] (see [12] for details). Each individual HOG cell has a side length of 8 pixels, and these cell features are densely computed across several scales with a scale factor of $1/\sqrt{2}$. We combine 2×2 adjacent cell features into a block feature yielding a single 124-dimensional local image feature that can be quantized into a visual HOG word. Each block is formed by computing the histograms for the individual cells first and then aggregating the cell histograms of blocks. Blocks are overlapping, as a new block starts at every HOG cell.

The description of the image content by HOG blocks is carried out analogous to that by SIFT features. HOG block features of an image are quantized into 10,000 discrete visual words using a flat visual vocabulary created with k-means clustering. The computed term-document vectors then serve as regular input to the topic models.

Note although HOG block features and SIFT features are on one side very much alike as both are effectively histograms of oriented gradients, they are on the other side also quite different with respect to the strictness with which they encode the spatial pattern of a local image region. SIFT encodes the spatial layout of gradients within a rigid 4×4 spatial grid, while in our case HOG employs a 2×2 spatial grid. Moreover, the gradients of each HOG cell are normalized by the gradient energies of surrounding cells. As a result SIFT is often used to identify patterns of specific objects such as of a specific landmark, a specific painting, etc. In contrast, HOG is usually used to identify object categories such bikes, people, cars, table, and a like.

Figures 3 and 4 show several examples of image patches that are described by the same visual words of SIFT features and HOG block features, respectively. Each row pair depicts sample patches of a different specific visual word.

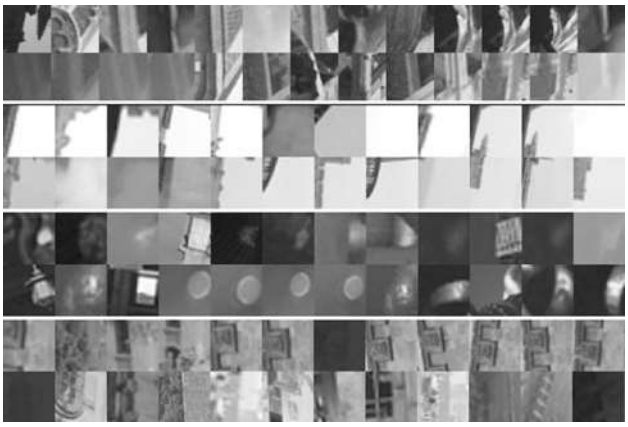

Fig. 3 Sample patches associated with four different visual word clusters of SIFT features derived from a vocabulary of 10,000 visual words

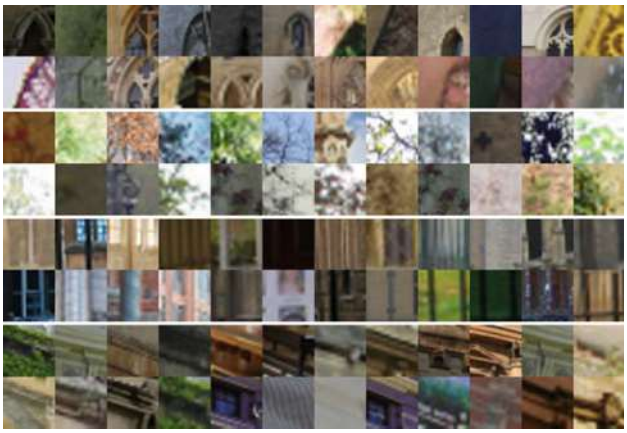

Fig. 4 Sample patches associated with four different visual word clusters of HOG block features derived from a vocabulary of 10,000 visual words. Note although HOG features are computed from color images, they effectively behave like grayscale features. Also they are not rotation invariant

3.5 Tag-based pLSA-model

Besides the visual description of an image we also consider tags as an additional modality. Tags are free-text annotations provided by the image authors or image owners. A tag can be single word as well as a phrase or a sentence. While Flickr stores the original form of an annotation such as “Golden Gate Bridge” in (here three) separate words, it further provides a generated raw tag like “goldengatebridge” that directly encodes the relationship of a particular word combination. In this work we treat each of these generated raw tags of the image annotations as one single word disregarding if it is a natural word or an artifically generated one. Thus, in the following the term *tag* denotes a single word derived from the raw tags and is used interchangeably with “word” and “term”.

As we use Flickr images to evaluate our multilayer multimodal pLSA model, it is important to note that these tags

Table 1 The vocabulary size before and after each filtering step. T_{minOcc} has been set to 1000 occurrences and T_{minUsers} has been set to 500 users

Number of images	10080251
Number of images with tags	9109593 (90.4%)
Number of Flickr users	852697
Vocabulary size after filtering step	
Number of all tags (unfiltered)	1691336
Removal of tags with length less than 2	1690029
Removal of tags that occur in less than T_{minOcc} images	6681
Removal of tags that contain numbers	6500
Removal of stop words	6467
Removal of tags used by less than T_{minUsers} different Flickr users	3158
Final vocabulary size	3158
#Images with tags within vocabulary	8803834 (87.3%)
Vocabulary words present in Wordnet	2483 (78.6%)

reflect the photographer/author's personal view with respect to the uploaded image. Thus, in contrast to carefully annotated image databases traditionally used for learning combined image and tag models [1], these image tags from Flickr are in many cases subjective, ambiguous, and do not necessarily describe the image content shown [20,22]. This makes it difficult to use the tags directly for retrieval purposes and thus some preprocessing is required. Even worse, some images do not have tags at all. In fact about 13% of all Flickr images lack annotations. In this case, textual information is not available for retrieval and a fallback strategy is needed. This underlines the importance of using a multimodal approach when exploiting user-generated content for image retrieval.

First a finite vocabulary needs to be defined, before a pLSA model can be applied to tags. Building the vocabulary starts with listing all tags that have been used more than T_{minOcc} times and by at least T_{minUsers} different users. This heuristics enforces that all rarely used tags are neglected. Note that a tag is also rarely used if only a few users have used it independent of the actually count. We further filter the list by discarding all tags that contain numbers. Table 1 shows the vocabulary sizes before and after filtering the available tags.

Once the tag vocabulary is defined, a co-occurrence table (i.e. a the term-document matrix) is built by counting the tag occurrences for each image. On average for annotated images the number of tags per images in our database is 7.7 (not counting tag-free images). For some images, however, the number of tags is unreasonably large as users have labeled images with whole sentences or phrases.

In our previous work [21], we used Wordnet [11] to expand the available image annotations. Wordnet is a lexical database of English that provides access to links and relationships between words. For each image we queried Wordnet for the semantic parents of the tags specified by the author. However, Wordnet is limited to English, and more than 20% of the words in our final vocabulary are not part of Wordnet (see Table 1). This may be caused by the use of different languages, slang words and abbreviations for annotations as well as the generation of raw tags that describe a specific location or scene. However, these annotations may carry very specific and meaningful information for correct retrieval. Therefore we do not restrict the annotations to plain English words. As the automatic expansion of textual words e.g. with hypernyms may also introduce additional noise to the annotations, we do not use Wordnet throughout this work and focus on the plain annotations provided by the image uploaders themselves.

In our experiments, we set the thresholds for the minimum number of occurrences $T_{\text{minOcc}} = 1000$ and for the minimum number of distinct users $T_{\text{minUsers}} = 500$ resulting in a vocabulary size of 3158 words. A larger tag vocabulary would be beneficial for a retrieval that is based solely on tags or other textual information. However, the training of the pLSA model is performed by sampling a subset of the whole database as training set (in this work 10,000 images). Thus, tags that do not occur within the set of training documents are not used for learning the pLSA model. In other words, tags that should be handled by the topic model need to be sufficiently frequent across all images in order to be included when (randomly) sampling the training set. This is the reason, why we chose this relatively small vocabulary for tags.

4 Multilayer multimodal pLSA

4.1 Motivation and model

In recent years, pLSA has been applied successfully to unimodal data such as text [16], image tags [24], or visual

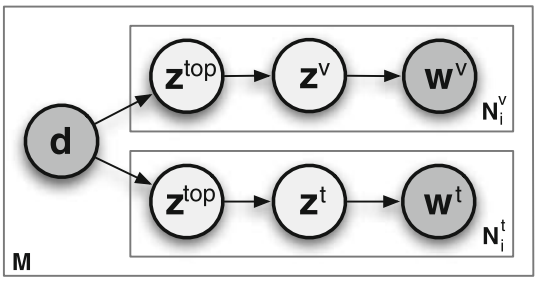

Fig. 5 The new multilayer multimodel pLSA model illustrated by combining two modalities

words [19]. However, combining two modes such as visual words and image tags is challenging. The obvious approach of simply concatenating the two associated term-document matrices $N_{M\times N^v}$ and $N_{M\times N^t}$ into $N_{M\times(N^v+N^t)}$ and then applying standard pLSA usually does not lead to the desired retrieval improvements. One reason is the difference in the order of magnitude with which words occur in the respective mode. For instance, a few thousand to 10,000 features per image are usually computed from images that are resized to having roughly the same number of dense samples while preserving the image's aspect ratio. In contrast, most images are annotated with fewer than 20 tags. Compensating between the differences in the order of the magnitude by some kind of normalization is possible, but will require a lot of testing to determine an appropriate weighting factor between the different modes since the actual importance of each mode must also be taken into account. Another reason may be the difference in the size of the respective vocabularies. In contrast, a well-founded mathematical approach with top-level topics will solve this issue effectively and efficiently. Some empirical evidence for these claims will be given in Sect. 5.

Our basic idea is to apply pLSA in a first step to each mode separately, and in a second step concatenate the derived topic vectors of each mode to learn another pLSA on top of that (see Fig. 7). While we describe this layering of multiple pLSAs only for two leaf-pLSAs and a node pLSA, it is obvious that the proposed pLSA layering can be extended to more than two layers and applied to more than just two leaf-pLSAs.

The smallest possible multilayer multimodal pLSA model (mm-pLSA) consisting of two modes with their respective observable word occurrences and hidden topics as well as a single top-level of hidden aspects is graphically depicted in Fig. 5. Every word of mode x (here: $x \in \{v, t\}$ with v standing for *visual* and t for *text*) occurring in document d_i is generated by an unobservable document model:

- Pick a document d_i with prior probability $P(d_i)$
- For each visual word in the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{\text{top}}|d_i)$
  - Select a visual topic z_k^v with probability $P(z_k^v|z_l^{\text{top}})$
  - Generate a visual word w_m^v with probability $P(w_m^v|z_k^v)$
- For each tag associated with the document:
  - Select a latent top-level concept z_l^{top} with probability $P(z_l^{\text{top}}|d_i)$
  - Select a tag topic z_p^t with probability $P(z_p^t|z_l^{\text{top}})$
  - Generate a tag w_n^t with probability $P(w_n^t|z_p^t)$

Thus, the probability of observing a visual word w_m^v or a tag w_n^t in document d_i is

$$P(d_i, w_m^v) = \sum_{l=1}^{L}\sum_{k=1}^{K} P(d_i)P(z_l^{\text{top}}|d_i)P(z_k^v|z_l^{\text{top}})P(w_m^v|z_k^v) \tag{5}$$

$$P(d_i, w_n^t) = \sum_{l=1}^{L}\sum_{p=1}^{P} P(d_i)P(z_l^{\text{top}}|d_i)P(z_p^t|z_l^{\text{top}})P(w_n^t|z_p^t). \tag{6}$$

An important aspect of this model is that every image consists of one or more part aspects in each mode, which in turn are combined to one or more higher-level aspects. This is very natural, since images consist of multiple objects parts and multiple objects. The multilayer multimodal pLSA can model this fact effectively—much better than a single layer pLSA. Furthermore, this model is in better correspondence with current belief to model the brain as a hierarchical recurrent network [14].

4.2 Training and inference

Given our word generation model (see Fig. 5) with its implicit independence assumption between generated words, the likelihood L of observing our database consisting of the observed pairs (d_i, w_m^v) and (d_i, w_n^t) from both modes is given by

$$L = \prod_{i=1}^{M}\left[\prod_{m=1}^{N^v} P(d_i, w_m^v)^{n(d_i, w_m^v)} \prod_{n=1}^{N^t} P(d_i, w_n^t)^{n(d_i, w_n^t)}\right]. \tag{7}$$

Taking the log to determine the log-likelihood l of the database

$$l = \sum_{i=1}^{M}\left[\sum_{m=1}^{N^v} n(d_i, w_m^v)\log P(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t)\log P(d_i, w_n^t)\right] \tag{8}$$

and plugging Eqs. (5) and (6) in to Eq. (8), it becomes apparent that there is a double sum inside of both *log*s making direct maximization with respect to the unknown probability distributions difficult. Therefore, we learn the unobservable probabilities distribution $P(z_l^{\text{top}}|d_i)$, $P(z_k^v|z_l^{\text{top}})$, $P(z_p^t|z_l^{\text{top}})$, $P(w_m^v|z_k^v)$ and $P(w_n^t|z_p^t)$ from the data using the EM-Algorithm [8]. Introducing the indicator variables

$$\triangle c_{lk} = \begin{cases} 1 & \text{if the pair } (d_i, w_m^v) \text{ was generated} \\ & \quad \text{by } z_l^{\text{top}} \text{ and } z_k^v \\ 0 & \text{otherwise} \end{cases}$$

$$\triangle d_{lp} = \begin{cases} 1 & \text{if the pair } (d_i, w_p^t) \text{ was generated} \\ & \quad \text{by } z_l^{\text{top}} \text{ and } z_p^t \\ 0 & \text{otherwise} \end{cases}$$

the complete data likelihood L_c, that is the data likelihood assuming that d_i, w_n^z, w_m^v, $\triangle c_{lk}$, and $\triangle d_{lp}$ are observable, is given by

$$L_c = \prod_{i=1}^{M} \left[\prod_{m=1}^{N^v} P(d_i, w_m^v, \triangle c)^{n(d_i, w_m^v)} \prod_{n=1}^{N^t} P(d_i, w_n^t, \triangle d)^{n(d_i, w_n^t)} \right]$$

with

$$\triangle c = (\triangle c_{11}, \ldots, \triangle c_{1K}, \ldots, \triangle c_{LK}) \tag{9}$$

$$\triangle d = (\triangle d_{11}, \ldots, \triangle d_{1K}, \ldots, \triangle d_{LP}) \tag{10}$$

$$P(d_i, w_m^v, \triangle c) = \prod_{l=1}^{L} \prod_{k=1}^{K} P(d_i) P(z_l^{\text{top}}|d_i) P(z_k^v|z_l^{\text{top}}) P(w_m^v|z_k^v)^{\triangle c_{lk}} \tag{11}$$

$$P(d_i, w_n^t, \triangle d) = \prod_{l=1}^{L} \prod_{p=1}^{P} P(d_i) P(z_l^{\text{top}}|d_i) P(z_p^t|z_l^{\text{top}}) P(w_n^t|z_p^t)^{\triangle d_{lp}} \tag{12}$$

Unlike in Eq. (8), we now only have product terms in the complete likelihood L_c, thus its log-likelihood can easily be termined and maximized,[1] resulting in the following expectation (E-step) and maximization (M-step) solution:

E-Step:

We estimate the unknown indicator variables $\triangle c_{lk}$ conditioned on the observable variables d_i and w_m^v by computing their expected value:

$$\begin{aligned} c_{lk}^{im} &:= E(\triangle c_{lk}|d_i, w_m^v) \\ &= P(\triangle c_{lk} = 1|d_i, w_m^v) \cdot 1 + P(\triangle c_{lk} = 0|d_i, w_m^v) \cdot 0 \\ &= P(\triangle c_{lk} = 1|d_i, w_m^v) \cdot 1 \\ &= \frac{P(d_i, w_m^v, \triangle c_{lk} = 1)}{P(d_i, w_m^v)} \\ &= \frac{P(d_i) P(z_l^{\text{top}}|d_i) P(z_k^v|z_l^{\text{top}}) P(w_m^v|z_k^v)}{\sum_{l=1}^{L} \sum_{k=1}^{K} P(d_i) P(z_l^{\text{top}}|d_i) P(z_k^v|z_l^{\text{top}}) P(w_m^v|z_k^v)}. \end{aligned} \tag{13}$$

Analogously, we estimate the unknown indicator variables $\triangle d_{lp}$ conditioned on the observable variables d_i and w_n^t by computing their expected value:

$$\begin{aligned} d_{lp}^{in} &:= E(\triangle d_{lp}|d_i, w_n^t) \\ &= \frac{P(d_i) P(z_l^{\text{top}}|d_i) P(z_p^t|z_l^{\text{top}}) P(w_n^t|z_p^t)}{\sum_{l=1}^{L} \sum_{k=1}^{K} P(d_i) P(z_l^{\text{top}}|d_i) P(z_p^t|z_l^{\text{top}}) P(w_n^t|z_p^t)} \end{aligned} \tag{14}$$

M-Step:

For legibility of the M-step estimates, we set

$$\gamma_{lk}^{im} := n(d_i, w_m^v) c_{lk}^{im} \tag{15}$$

$$\delta_{lp}^{in} := n(d_i, w_n^t) d_{lp}^{in} \tag{16}$$

which is the expected probability of observing a pair (d_i, w_m^v) multiplied with the actual number of occurrences and get:

$$P(d_i)^{\text{new}} = \frac{\sum_{m=1}^{N^v} n(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t)}{\sum_{i=1}^{M} \left(\sum_{m=1}^{N^v} n(d_i, w_m^v) + \sum_{n=1}^{N^t} n(d_i, w_n^t) \right)} \tag{17}$$

$$P(z_l^{\text{top}}|d_i)^{\text{new}} = \frac{\sum_{m=1}^{N^v} \sum_{k=1}^{K} \gamma_{lk}^{im} + \sum_{n=1}^{N^t} \sum_{p=1}^{P} \delta_{lp}^{in}}{\sum_{l=1}^{L} \left(\sum_{m=1}^{N^v} \sum_{k=1}^{K} \gamma_{lk}^{im} + \sum_{n=1}^{N^t} \sum_{p=1}^{P} \delta_{lp}^{in} \right)} \tag{18}$$

$$P(z_k^v|z_l^{\text{top}})^{\text{new}} = \frac{\sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im} + \sum_{p=1}^{P} \sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}} \tag{19}$$

$$P(z_p^t|z_l^{\text{top}})^{\text{new}} = \frac{\sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{m=1}^{N^v} \gamma_{lk}^{im} + \sum_{p=1}^{P} \sum_{i=1}^{M} \sum_{n=1}^{N^t} \delta_{lp}^{in}} \tag{20}$$

$$P(w_m^v|z_k^v)^{\text{new}} = \frac{\sum_{i=1}^{M} \sum_{l=1}^{L} \gamma_{lk}^{im}}{\sum_{m=1}^{N^v} \sum_{i=1}^{M} \sum_{l=1}^{L} \gamma_{lk}^{im}} \tag{21}$$

$$P(w_n^t|z_p^t)^{\text{new}} = \frac{\sum_{i=1}^{M} \sum_{l=1}^{L} \delta_{lp}^{in}}{\sum_{n=1}^{N^t} \sum_{i=1}^{M} \sum_{l=1}^{L} \delta_{lp}^{in}} \tag{22}$$

Clearly, Eq. (17) is constant across all iterations and must not be recomputed.

[1] A complete derivation of the EM-update equation for this multilayer multimodel pLSA model can be found at http://www.multimedia-computing.de/wiki/mm-pLSA

Given a new test image d_{test}, we estimate the top-level aspect probabilities $P(z_l^{\text{top}}|d_{\text{test}})$ with the same E-step equations as for learning and Eq. (18) for $P(z_l^{\text{top}}|d_{\text{test}})$ as the M-step. The probabilities of $P(z_k^v|z_l^{\text{top}})$, $P(z_p^t|z_l^{\text{top}})$, $P(w_m^v|z_k^v)$ and $P(w_n^t|z_p^t)$ have been learned from the corpus and are kept constant during inference.

Remark 1 Normalization Before starting the mm-pLSA the document vectors of different modalities, i.e. the entries $n(d_i, w_m^v)$ and $n(d_i, w_n^t)$ should be normalized to equal scale, e.g. such that the sums over each modality separately are equal. This is crucial if one modality has document vectors on a very different scale than the other modality, e.g. compare the highly populated histograms of visual features to very sparse tag histograms. In that case the mm-pLSA on unnormalized feature histograms is dominated by the visual domain and the probabilities $P(z_p^t|z_l^{\text{top}})$ would be close to zero. Note that this normalization does not mean that e.g. visual and textual modality have the same weight within the mm-pLSA as the constraint for the conditional probabilities of the subtopics given the supertopics is given by

$$\sum_{k=1}^{K} P(z_k^v|z_l^{top}) + \sum_{p=1}^{P} P(z_p^t|z_l^{top}) = 1$$

In fact we noticed that the mm-pLSA on SIFT features and tags determines a higher weight for the textual domain. See Sect. 5.4 for further details.

Remark 2 Training The training itself must only consider documents that have non-zero document vectors for both domains. With missing co-occurrences across the modalities the model training is useless. However, the inference still is able to derive a topic distribution even if one modality (e.g. annotations) is not available for an image.

Remark 3 Training Furthermore the training procedure should sample training documents such that basically all visual and textual aspects that appear in the database are also present in the training set. However the number of images for a certain class or category may vary. Therefore we pseudo-randomly pick training samples by selecting documents at certain intervals from the whole list of documents starting at a random offset. This guarantees that the whole database is used when drawing samples disregarding the actual layout and order. Training documents of a certain category are drawn with a probability corresponding to its size.

4.3 Fast initialization

More complicated probabilistic models always come with an explosion in required training time. This issue is becoming more severe, the more layers and the more pLSAs are aggregated into higher-level pLSAs. Thus, we suggest to compute a decent initial estimation of the conditional probabilities in a strictly stepwise forward procedure (see Fig. 7) as proposed in [27].

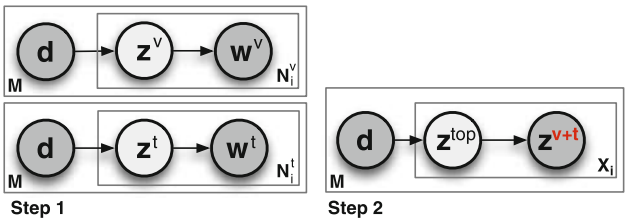

Fig. 6 The fast initialization of the multilayer multimodal pLSA model computed in two separate steps

For the smallest two-leaf high-level aspect model this procedure first computes an independent pLSA for each mode on the lowest level. The aspects are only linked through the documents, ie., the same images (see Step 1 in Fig. 6). Next the computed aspect of all modes are taken as the observed words at the next higher level (see Step 2 in Fig. 6). This procedure can continue until the top-level aspect vector is learned. The final representation, the top-level aspect distribution for each document, describes each image as a "distribution over topic distributions" and thereby fuses the visual pLSA model and the tag pLSA model. An overview of such an image retrieval system based on this idea is shown in Fig. 7.

As we will show in the experimental results, this fast initialisation already produces a decent model. It can be further be improved by appying the EM-algorithm as stated in Sect. 4.2 to the complete model after initializing it with the strictly forward computed solution. This will further improve the solution.

Figure 8 shows the development of the complete data log-likelihood along the increasing number of iterations. One can observe that the mm-pLSA training converges much faster when initialized with the former multimodal standard pLSA solution over random initialization.

5 Experimental evaluation

5.1 Setup

For each of the visual features (SIFT, HOG) and the tag features we learned a 50-topic pLSA model. The fast initialization of the mm-pLSA mapped the two 50-dimensional image representations computed by the two base models (based on visual features and tags) to a multimodal topic distribution over 50 "super" topics. The randomly initialized mm-pLSA and its optimized version with the general mm-pLSA learning algorithm directly computed a model with 50 topics. The number of iterations used during training and inference varied. All models were computed using 500 iterations,

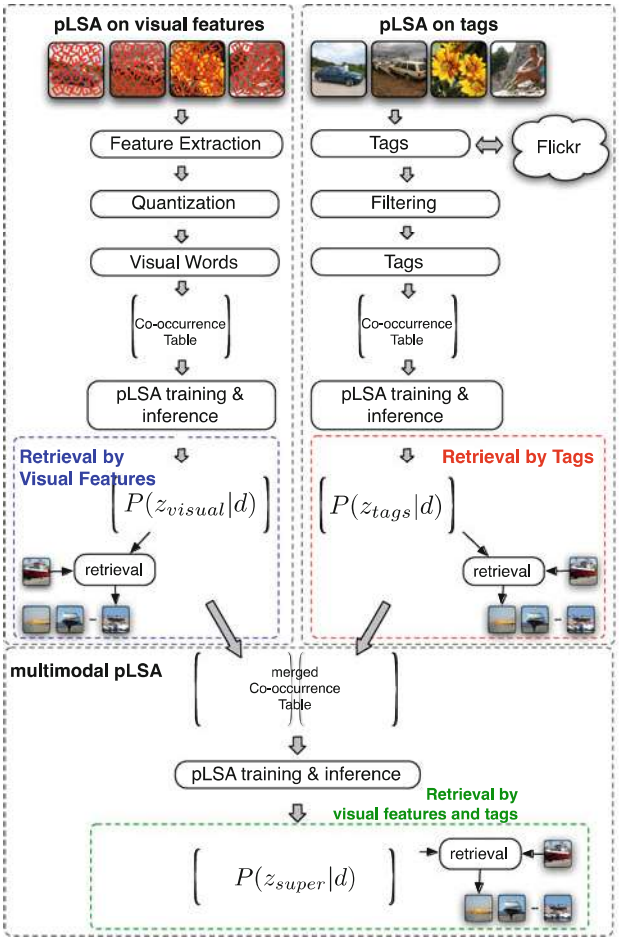

Fig. 7 Schematic overview of the retrieval system based on our fast initialization strategy. Given the fast initialization the subsequent full mm-pLSA optimizes all three steps at once

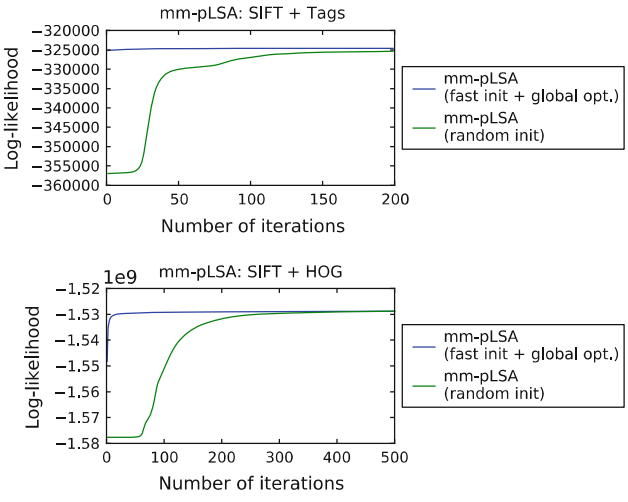

Fig. 8 Log-likelihood over training data when learning the mm-pLSA model. The mm-pLSA initialized by the strictly stepwise forward multimodal pLSA converges much faster than the model starting from a random initialization. The *upper image* shows the log-likelihood when the mm-pLSA is applied for SIFT features and tags, the *lower image* shows the log-likelihood for SIFT and HOG block features

except the mm-pLSA with the fast initialization method. In this case the model was computed using 50 iterations, since we already had a good starting point. Each pLSA model, independent of whether a conventional unimodal or a multilevel multimodal pLSA model was trained with 10,000 images.

The only probability distribution computed during inference was the probability distribution $P(z_l^{\text{top}}|d_i)$ of the top-level topics given the document. Therefore the EM-algorithm converged faster than during training and the number of iterations was reduced. For the inference of these topic distributions we used 200 iterations with the visual-based pLSA, the tag-based pLSA, the concatenated topic-based pLSA, the fast initialization of the mm-pLSA. 50 iterations were used for the inference of the mm-pLSA models both on visual features and tags and for all modes (either randomly initialized or using the fast initialization).

We evaluated all the systems in a query-by-example task and evaluated the results by a user study with 9 users. 80 query images were selected and the L_1 distance was used to find the most similar images. The query images are associated with their original tags, while we only kept queries where the original annotation roughly correspond to the image content. The participants were asked to rate the 19 closest results to each of our query images. Note that we always showed the images without their associated tags as we evaluated a query-by-image-example system. We used the following scoring to get a quantitative performance measure: An image considered being similar received 1 point, an image considered somewhat similar received 0.5 points. All other images got 0 points. A mean score was calculated for each user; the mean over all users' means yielded the final score of the system being evaluated. Two example queries and the topmost retrieved images are shown in Fig. 13.

As we also evaluate one system that is based solely on tags, it happens that there are several hundreds up to thousands of images that have the same distance to the query image. This is due to the fact that images annotated with the same words will yield the same topic distribution disregarding the image content. For an unbiased evaluation the images in the result list need to be sorted by ascending distance (as usual) with an additional randomization step for images with equal distances. That is, images with equal distance to the query are randomized in their order while the ascending order of distances is still maintained for the whole list. This procedure eliminates any bias introduced by the order, in which similar images are found when scanning through the database (Table 2).

We further impose two additional constraints:

– Any retrieved image from the same Flickr user who uploaded the query image will be ignored.

Table 2 Example categories in Flickr-10M

Landmarks	Scenes	Objects
Abu Simbel, Allianz arena, Angel falls, Arc de Triomphe, Church of Saviour Blood, Ayers Rock, Banaue Rice Terrace, Basilica de Notre Dame, Berlin Wall, Big Ben, Bilbao Gugenheim Museum, Biosphere Montreal, ...	Beach, Carnival, Christmas, City, Desert, Forest, Portrait, Street, Sunset, Wedding	Aircraft, Bicycle, Bird, Boat, Bottle, Building, Bus, Butterfly, Car, Cat, Chair, Cow, Dog, Fish, Flower, Horse, ...
Activities	National Parks	Stars
Aikido, Archery, Arm wrestling, Ax throwing, Badminton, Ballett, Baseball, Basketball, Belly dance, Billards, BMX, Bowling, Boxing, ...	Abel Tasman, Acadia, Addo Elephant, Algonquin, Ayuittuq, Bandhavgarh, Banff, Bromo Tenger, Cuc Phuong, Gran Paradiso, ...	Alice Cooper, Angenlina Jolie, Ashley Olsen, Audey Hepburn, Barack Obama, Ben Stiller, Bill Clinton, Bill gates, Bono, Brad Pitt, Britney Spears, Bruce Willis, Bryan Adams, ...
Total number of images (without duplicates)		10,080,251

The full list is available at http://www.multimedia-computing.de/wiki/Flickr-10M

– Any Flickr user may only contribute a single image to the result set. This is the one with the smallest distance, other retrieved images of that specific user will be ignored.

These restrictions minimize the impact of image series uploaded by a single user to the evaluation.

5.2 Dataset

We have created a new publicly available dataset called "Flickr-10M"[2] to evaluate the proposed retrieval methodology on a large real-world image database. This data set consists of 10 million images downloaded from Flickr.

We aimed to make this dataset as diverse as possible to allow the evaluation of greatly varying retrieval approaches. Therefore we collected images that were annotated with specific tags, which indicate a variety of landmarks, scenes, cities, stars as well as objects. Geotags were explicitly not used to download images for two reasons: In most cases, the number of images that actually have been geo-tagged is very small even for popular landmarks. Furthermore many landmarks are photographed from the far distance. In that case the geo-tagged location may be far from the position of the landmark itself. Also, for many categories like cities or national parks geotags are relatively meaningless despite narrowing down the number of available images. Therefore, we focused on tags and image descriptions. In cases a certain category did not yield a sufficient number of images (e.g. several thousands) we performed a full-text search for the query term in the image description to select the downloaded images (See Table 2 for examples).

This size of the dataset is beyond most datasets targeting a specific domain like scenes (e.g. SUN database [31]), objects (e.g. PASCAL VOC [10]), or landmarks (e.g. Oxbuild [26]). It is comparable in its size to Imagenet [9] and orders of magnitudes bigger than datasets that were previously used for image retrieval evaluations like Oxbuild or Corel.

This dataset consists of JPEG images with their associated metadata. This includes tags, titles, descriptions, and other user-generated content as well as other information stored with the photos (e.g. EXIF data if available). There are 852,697 different Flickr users that contribute at least one photo to our dataset. In total there are more than 300 different categories yielding a total of 10,080,251 images.

The database has not been cleaned or post-processed. Thus, it includes all kinds of content, e.g. from high-quality to low-quality photographs with and without annotations in all kinds of languages. In short, we believe this database is a representative sample of the real data that is uploaded and shared on community websites and social networks on a daily basis.

5.3 Results

First, we evaluate the fusion of the visual domain (represented by SIFT features) with the image annotations. The results of this experiment are shown in Fig. 9. The first two experiments measure the performance of the systems based solely on visual features or tags and are labeled "pLSA on SIFT" and "pLSA on tags", respectively. "Concatenated pLSA" denotes the model computed from merging the words from the visual domain as well as the tag domain into a single feature vector. The straight-forward approach of applying a third pLSA model on top of the two base models is termed "mm-pLSA (fast init only)", while the mm-pLSA that is initialized randomly or with the outcome of the fast initialization is denoted as "mm-pLSA (random init)" or "mm-pLSA (fast init)", respectively.

It can be seen that the system relying solely on tags performs worse than the system relying solely on visual features. This is somewhat unexpected as in previous work tags were

[2] The dataset and additional material are available at http://www.multimedia-computing.de/wiki/Flickr-10M

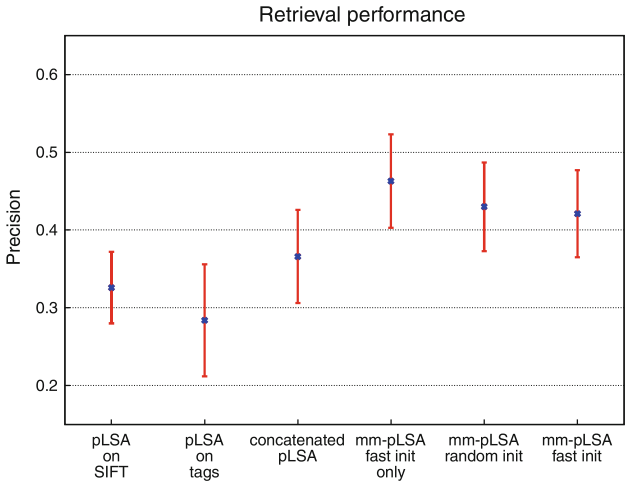

Fig. 9 Scores for our different retrieval systems based on SIFT features and tags. *Vertical bars* mark the standard deviation between the users' means

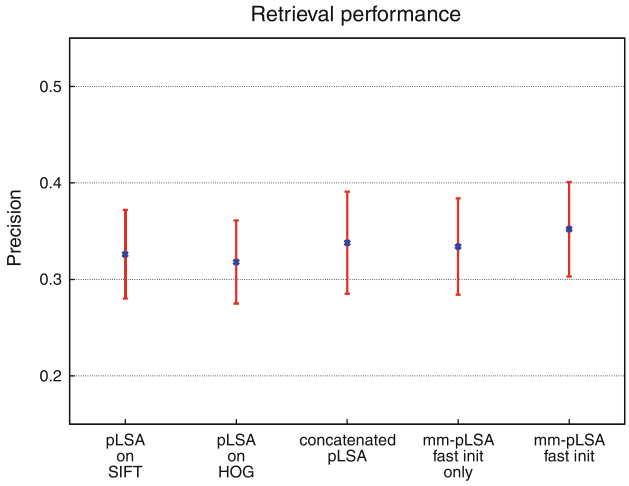

Fig. 10 Scores for our different retrieval systems based on SIFT and HOG features. *Vertical bars* mark the standard deviation between the users' means

shown to outperform the visual features alone (see [21] for details). The third system, aiming to fuse the modalities by simply concatenating the (normalized) occurrence counts, performs better than the unimodal systems but worse than than any mm-pLSA model.

Both mm-pLSA models with fast initialization only and with optimizing the already good initialization outperform the unimodal modals which confirms the expected superior performance of multimodal models. However, the mm-pLSA models with global optimization (either random initialization or fast initialization strategy) perform slightly worse than the model that only performs the fast initialization. This is unexpected and somewhat contradictory to previous works [21]. We suspect that the global optimization drifts too towards the textual domain. Given the poor performance of tags alone the overall performance then suffers. Another possible reason is that the global optimization is unable to optimize the solution from the fast initialization strategy any further. Figure 8 shows that the log-likelihood of that model does hardly increase. This may be caused by too much noise on image annotations or a too small number of training documents.

The randomly initialized mm-pLSA model performs worse than the mm-pLSA with fast initialization strategy. This is in line with our expectations: we expected a random initialized model to perform inferior to its well initialized counterpart. It should be noted that as the EM-algorithm already starts from a relatively good solution, the number of required training iterations is small. Therefore the training of the mm-pLSA with the fast initialization strategy is fast and effective.

In a second serious of experiments, we evaluate how the mm-pLSA can be used to fuse multiple features into a combined representation. In these experiments the two modalities that are evaluated are SIFT and HOG features. The results of the corresponding user studies are shown in Fig. 10. Similar to the previous experiments, the pLSA on the concatenated feature histograms does hardly improve over the better of the two modalities. This observation underlines the importance of hierarchical models even for assumed easy tasks such as multi-feature combination. Despite the close relation of these gradient-based features one can see that a stepwise combination of three pLSA models (termed "mm-pLSA fast init only") further improves the retrieval, but is slightly outperformed by the mm-pLSA model that performs a global optimization.

It remains subject of future research why the mm-pLSA model with fast initialization strategy and global optimization performs worse than expected on this data set but outperformed all other in previous work in the case where SIFT features and tags combined. A probably related issue is the inferior performance of the tag-based model. One possible solution may be to upscale the tag vocabulary in order to describe such huge data set more accurately. Another potential solution may be to also include the provided textual image description of Flickr images rather than tags alone.

5.4 Discussion

For further insights we visualize the conditional probabilities of the modality-specific "subtopics" given the "supertopics" ($P(z_k^v|z_l^{\text{top}})$ and $P(z_p^t|z_l^{\text{top}})$) of the mm-pLSA training. We chose the mm-pLSA with fast initialization strategy and plot these probabilities as a matrix, where the actual probability value is mapped to a color ranging from dark black for 0 to bright white for 1. Each row l of such a matrix represents $P(z_k^v|z_l^{\text{top}})$ on the left half (split by the red line)

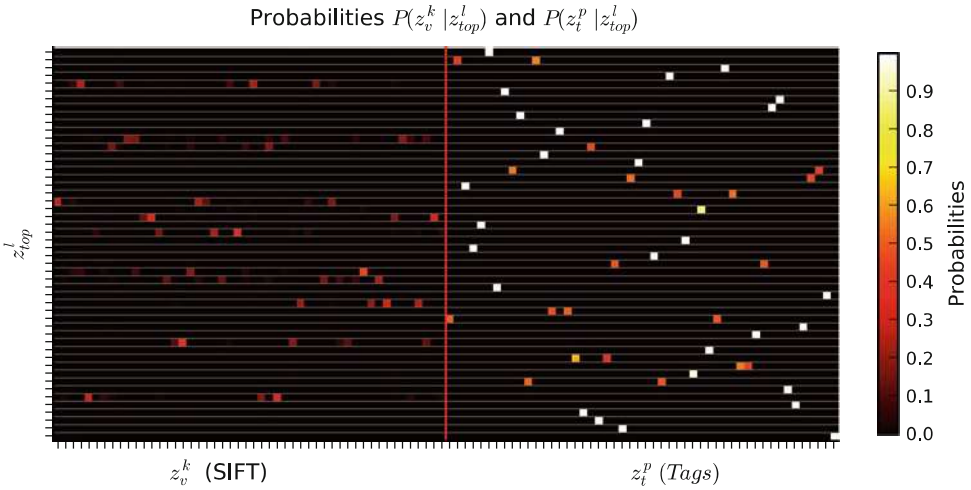

Fig. 11 Visualization of the matrix $P(subtopics|supertopics)$ for the mm-pLSA on SIFT features and tags. One row in this matrix denotes all conditional probabilities $P(z_k|supertopics)$ and $P(z_p|supertopics)$ summing to 1. The subtopics for the SIFT features are shown on the *left half*, the subtopics derived from tags on the *right half*. (Best viewed in color) (color figure online)

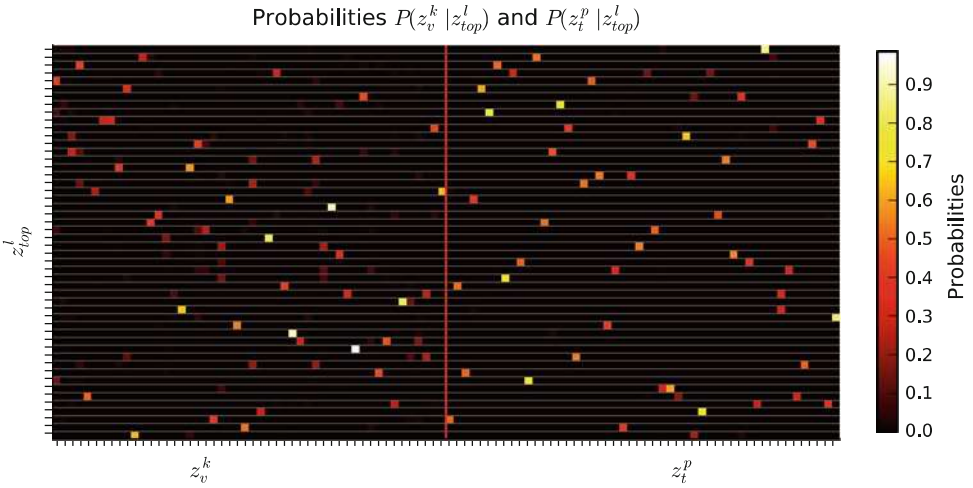

Fig. 12 Visualization of $P(subtopics|supertopics)$ for the mm-pLSA on SIFT and HOG features. One row in this matrix denotes all conditional probabilities $P(z_k|supertopics)$ and $P(z_p|supertopics)$ summing to 1. The subtopics for the SIFT features are shown on the *left half*, the subtopics derived from HOG features on the *right half*. (Best viewed in color) (color figure online)

and $P(z_p^t|z_l^{\text{top}})$ on the right half. The columns then enumerate the subtopics k and p correspondingly. Note that each row sums to 1. Therefore one can easily identify the present mixture of the modalities by looking at each row.

The conditional probabilities for SIFT features and tags are shown in Fig. 11. It can be seen that most entries with high probability value are present for tags only (right half of Fig. 11). The visual part (left half) has no peaks but is apparently less sparse. One can further observe that the entries in each row with a significant probability (the visible entries) are either on the visual or on the textual side, not on both. There is no direct correspondence between visual topics and textual topics. This means that each (super-) topic determined by the mm-pLSA basically acts as a kind of auto-selection mechanism for these two modalities. The mixture of visual and textual description is thereby achieved by representing each individual image by a mixture of such supertopics. These are in turn mutually exclusive on their subtopic representation, but the mixture of these describes both modalities.

This is different for the multi-feature model combining SIFT and HOG features. In Fig. 12, one can see that the supertopics represent a real mixture of subtopics from different modalities.

6 Conclusion

A very general scheme for multilayer multimodal probabilistic Latent Semantic Analysis has been proposed. It naturally extends the single-layer pLSA to the concept of layered or hierarchical topics—a natural way to describe an image composition. It also allows grasping concepts across different modalities. The proposed fast initialization technique makes the mm-pLSA very practical and computable. The

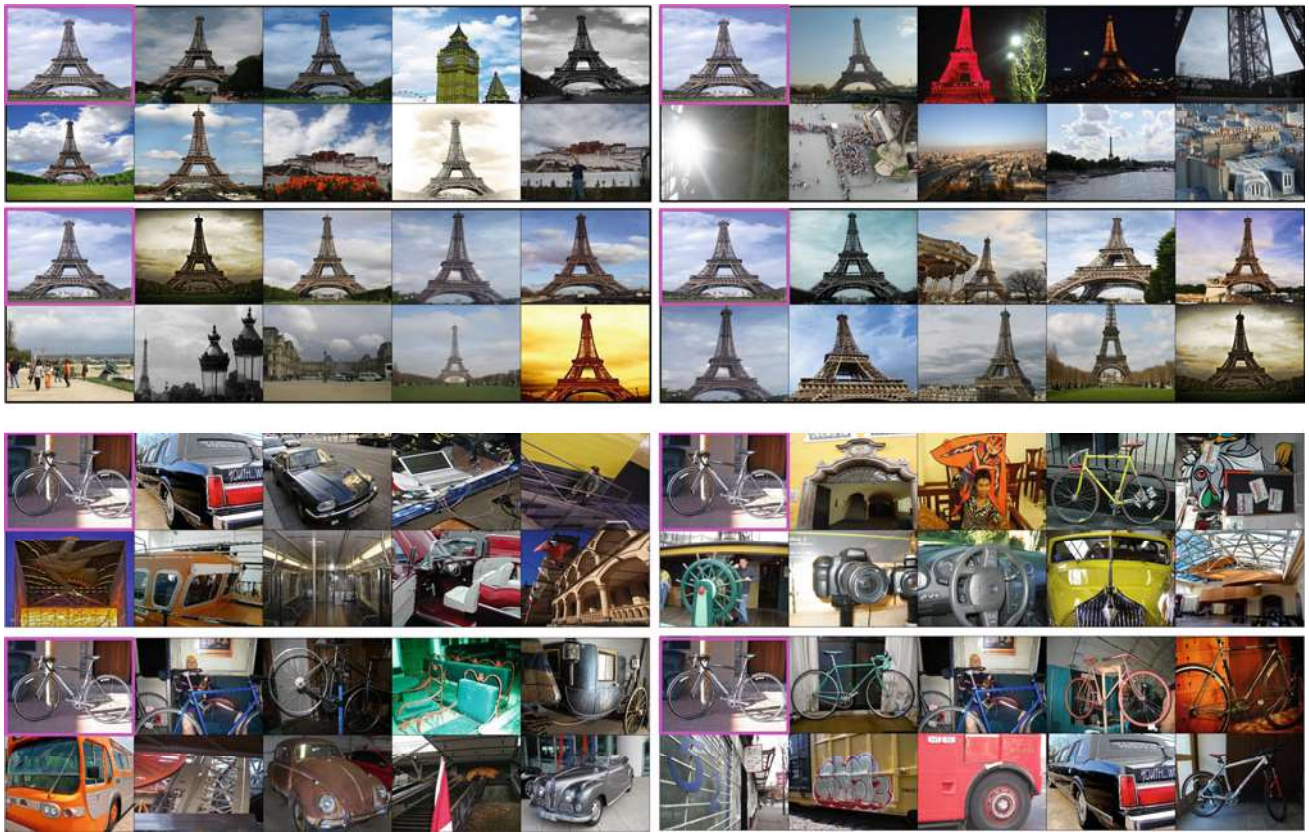

Fig. 13 Examples of retrieval results for the different approaches and two different queries. The query image is shown at the *top left corner* (*pink frame*) followed by the retrieved images. **Query: "Eiffel Tower"**: *Upper left* pLSA on SIFT features. *Upper right* pLSA on tags. *Lower left* mm-pLSA (the fast initialization only) on both SIFT and tags. *Lower right* mm-pLSA with fast init and global optimization on both SIFT and tags. **Query: "bike"**: *Upper left* pLSA on SIFT features. *Upper right* pLSA on HOG features. *Lower left* mm-pLSA (the fast initialization only) on both SIFT and HOG features. *Lower right* mm-pLSA with fast init and global optimization on both visual feature types

overall approach was evaluated in a query-by-example image retrieval scenario by users and outperformed unimodal pLSA significantly. The simple structure of two leaves, one node instance of such model was just an example and can be extended to full tree structures with more than two layers. Thus the mm-pLSA shows huge promise for future research (See Fig. 13 for example queries and the corresponding retrieval results).

Acknowledgments We thank Deutsche Forschungsgemeinschaft (DFG) for funding this project.

References

1. Barnard K, Duygulu P, Forsyth D, Blei DM, Hofmann T, Poggio T, Shawe-taylor J (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
2. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) SURF: speeded up robust features. Comput Vis Imag Underst 110(3):346–359
3. Berg AC, Berg TL, Malik J (2005) Shape matching and object recognition using low distortion correspondences. In: IEEE conference on computer vision and pattern recognition (CVPR'05), vol 1. Washington, DC, pp 26–33
4. Blei D, Lafferty J (2006) Correlated topic models. In: Advances in neural information processing systems, vol 18, pp 147–154
5. Blei DM, Jordan MI (2003) Modeling annotated data. In: ACM SIGIR conference on research and development in information retrieval (SIGIR'03), pp 127–134
6. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
7. Bosch A, Zisserman A, Muñoz X (2006) Scene classification via pLSA. Eur Confer Comput Vis (ECCV'06) 3954:517–530
8. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39(1):1–38
9. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR'09)
10. Everingham M, Van Gool L, Williams C, Winn J, Zisserman A (2009) The pascal visual object classes (VOC) challenge. Int J Comput Vis (IJCV'04) 88(2):303–338
11. Fellbaum C (1998) WordNet: an electronic lexical database. MIT Press, Cambridge
12. Felzenszwalb P, Girshick R, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell (PAMI'10), 32(9)
13. Greif T, Hörster E, Lienhart R (2008) Correlated topic models for image retrieval. Technical Report TR2008-09, University of Augsburg

14. Hawkins J, Blakeslee S (2004) On intelligence. Times Books, New York
15. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
16. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1–2):177–196
17. Hörster E, Lienhart R (2008) Deep networks for image retrieval on large-scale databases. In: ACM international conference on multimedia (MM'08), New York, pp 643–646
18. Hörster E, Lienhart R, Slaney M (2007) Image retrieval on large-scale image databases. In: ACM international conference on content-based image and video retrieval (CIVR'07), pp 17–24
19. Hörster E, Lienhart R, Slaney M (2008) Continuous visual vocabulary models for pL-based scene recognition. In: ACM international conference on content-based image and video retrieval (CIVR'08), New York, pp 319–328
20. Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T (2007) How flickr helps us make sense of the world: context and content in community-contributed media collections. In: ACM international conference on multimedia (MM'07), New York, pp 631–640
21. Lienhart R, Romberg S, Hörster E (2009) Multilayer pLSA for multimodal image retrieval (CIVR'09). In: ACM international conference on image and video retrieval, vol 14
22. Lienhart R, Slaney M (2007) pLSA on large scale image databases. IEEE Int Confer Acoust Speech Signal Process (ICASSP'07) IV:1217–1220
23. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis (IJCV'04) 60(2):91–110
24. Monay F, Gatica-Perez D (2004) pLSA-based image auto-annotation: constraining the latent space. In: ACM international conference on multimedia (MM'04), New York, pp 348–351
25. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. IEEE Confer Comput Vis Pattern Recogn (CVPR'06) 2:2161–2168
26. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. IEEE Confer Comput Vis Pattern Recogn (CVPR'07) 3613:1575–1589
27. Romberg S, Horster E, Lienhart R (2009) Multimodal pLSA on visual features and tags. In: IEEE international conference on multimedia and expo (ICME'09), pp 414–417
28. Shechtman E, Irani M (2007) Matching local self-similarities across images and videos. In: IEEE conference on computer vision and pattern recognition (CVPR'07)
29. Sivic J, Russell BC, Zisserman A, Freeman WT, Efros AA (2008) Unsupervised discovery of visual object class hierarchies. In: IEEE conference on computer vision and pattern recognition (CVPR'08)
30. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: International conference on computer vision (ICCV'03)
31. Xiao J, Hays J, Ehinger K, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: IEEE conference on computer vision and pattern recognition (CVPR'10)
32. Zhang L, Wang X-j (2011) Multi-Feature pLSA for combining visual features in image annotation. In: ACM international conference on multimedia (MM'11), Scottsdale, Arizona, pp 1513–1516